RocksDB on Zoned NVMe™ SSDs

- Optimal data placement
- Minimal Device Write amplification
- No drive garbage collection
- Longer life
- Higher throughput
- Lower latency

RocksDB

ZNS NVMe SSD Development Platform
Agenda

- Zoned Namespaces 101
- Adapting RocksDB for Zoned SSDs
- Demo
- Results
- What’s next?
Zoned Namespaces 101
What are Zoned Block Devices?

The new paradigm in storage

- The storage device logical block addresses are divided into ranges of zones.
- Writes within a zone must be sequential.
- The zone must be erased before it can be rewritten.
Zoned Storage on SMR

- SMR (Shingled Magnetic Recording)
  - Enables areal density growth
  - Shares flash access model
    - Erase before re-write

- Zoned Access
  - Zoned Block I/F standardized in INCITS
    - Zoned Block Commands (ZBC): SAS
    - Zoned ATA Commands (ZAC): SATA
  - Host/Device cooperate to optimize RMW aspect of SMR by enforcing sequential writes and enabling host FTL model
Ubiquitous Workloads

The cloud applies multiple workloads to a single SSD

SSDs write log-structured to the media that requires garbage collection

Multiplex data streams onto the same garbage collection units

Increases Write Amplification, Over-Provisioning and thereby Cost

Decreases throughput and latency predictability
Zones for Solid State Drives

Eliminate data streams multiplexing:
- **Significantly decreases** write amplification, over-provisioning and **thereby reduces cost**
- **Increases throughput and latency predictability**
ZNS: Synergies w/ ZAC/ZBC software ecosystem

- Device exposed as a Zoned Block Device (ZBD)
- Reuse existing work already done for ZAC/ZBC devices
- Existing ZBD-aware file systems & device mappers “just work”
  - Few additions to support to ZNS
- Integrates with file-systems and applications
  - RocksDB, Ceph, fio, libzbd, …
- ZAC/ZBC devices are already in production at technology adopters and a mature storage stack is available through the Linux® eco-system

* = Enhanced data paths for SMR/ZNS drives
Zoned Namespaces

- Ongoing Technical Proposal in the NVMe working group
- New Zoned Command Set – Inherits the NVM Command Set and adds zone support.
- Aligns to the existing host-managed models defined in the ZAC/ZBC specifications.
  - Note that it does not map 1:1. Beware of the details.
- Optimized for Solid State Drives
  - Zone Capacity
  - Zone Append
  - Zone Descriptors

![Diagram of Zoned Namespaces]
Host-Managed Zoned Block Devices

- Zone States
  - Empty, Implicitly Opened, Explicitly Opened, Closed, Full, Read Only, and Offline.
  - Changes state upon writes, zone management commands, and device resets.
- Zone Management
  - Open Zone, Close Zone, Finish Zone, and Reset Zone
- Zone Size & Zone Capacity
  - Zone Size is fixed
  - Zone Capacity is the writeable area within a zone

![Diagram of zone states and management](image)

Zone Size (e.g., 512MB)

Zone Start LBA

Zone Capacity (E.g., 500MB)

Zone X

Zone X - 1

Zone X + 1

Active Resources

Open Resources

- Empty (ZS0)
- Full (ZS1)
- Closed (ZS2)
- Implicitly Opened (ZS3)
- Explicitly Opened (ZS4)
- Read Only (ZS5)
- Offline (ZS6)

Zone Management Commands

- Open Zone
- Close Zone
- Finish Zone
- Reset Zone

Zone Capacity (E.g., 500MB)
Zoned Storage is a class of storage devices that enables host and storage devices to cooperate to achieve higher storage capacities, increased throughput, and lower latencies. The zoned storage interface is available through the SCSI Zoned Block Commands (ZBC) and Zoned Device ATA Command Set (ZAC) standards on Shingled Magnetic Recording (SMR) hard drives today and is also being adopted for NVMe Solid State Disks with the upcoming NVMe Zoned Namepaces (ZNS) standard.

QUICK START GUIDE

Get Started »

Linux Kernel Features

Linux kernel supports zoned storage devices through various I/O paths with different access characteristics, such as raw device access, file systems and device mapper targets.

View Details »

System Compliance Tests

Learn how to verify a system readiness for zoned storage devices and test hardware components compliance to standards with automated test suites.

View Details »
RocksDB on ZNS
RocksDB, a good fit for ZNS

- Persistent key-value store for fast storage environments
- Log-structured, flash friendly
- Customizable storage back ends
WA on a conventional SSD

Device Write Amplification Factor

Disk Space in Use

7 billion inserts followed by 5 billion overwrites

4X
Target: End-to-end-integration
Challenges

- Multiple, parallel files being written to
  - Map each file to a set of zones
- All writes must be sequential and ordered
  - Use direct I/O and the deadline scheduler
- Limits on number of open zones
  - Finish zones when done with writes
RocksDB on-disk data structures

Writeahead-log (WAL)

<KEY, VALUE>

WAL

Memtable
RocksDB on-disk data structures

Sorted string tables

- **Persisted In memory**
  - e.g. 64MB
  - e.g. 128MB
  - e.g. 1G
  - Grows 10X
  - e.g. 100G

- **In memory**
  - Append-only memtable
  - L0
    - Most Updated
    - e.g. 64MB
  - L1
    - e.g. 128MB
    - e.g. 1G
    - Grows 10X
  - Ln
    - Least Updated
    - e.g. 100G

- **Sync**
  - compaction
  - sstable

- **Hot Data**
- **Cold Data**
Mapping files to zones

- WAL File
- WAL File
- SST File
- SST File
- SST File
- SST File
Approach

- Files are mapped to zones
- Zone management through file management
  - Zones are allocated when creating a new file
  - Zones are released after file deletion
  - Zones can be rewritten after being reset

No device-side garbage collection
Demo (Random Insert Workload)
Results

Smart data placement

~1X device write amplification
3-6X WA’s measured on conventional drives (28% OP)

No on-drive garbage collection

20% increase in capacity
Compared to a conventional 28% OP SSD

20% TCO reduction
Increases lifetime/writes significantly
Conclusions

- Easy to leverage flash-friendly data placement
  - ZNS enables applications to become flash-optimal
- Zoned Block Device Software Eco-system already available
  - Libraries, tools, emulation
- Easy to integrate with existing storage stack
What’s next?

- Upstream support to RocksDB
- More ZNS end-to-end-integration:
  - Databases (LSM-based, logs, ...)
  - Filesystems (btrfs, ceph, ...)
- Cloud infrastructure
Thanks!

Western Digital and the Western Digital logo are registered trademarks or trademarks of Western Digital Corporation or its affiliates in the US and/or other countries. The NVMe word mark and NVM Express design mark are trademarks of NVM Express, Inc. All other marks are the property of their respective owners.